Journal of the
International Institute for Terminology Research
– IITF –

TERMINOLOGY SCIENCE & RESEARCH

Vol. 21 (2010)

Publisher:
Address:
Redaktion: International Institute for Terminology Research (IITF)

CONTENTS

Foreword

Volume 20 (2009) of the Journal Terminology Science and Research contained four of the papers presented at the panel session organized by the IITF within the framework of the XVIII LSP Symposium in Aarhus, August 2009. The present volume of TSR contains one of the remaining papers, that of Gianna Tarquini, and one additional article.

In accordance with the overall theme of the workshop, Conceptual representation in terminology across various semiotic systems and media, Tarquini's paper deals with the problem of how to treat non-verbal and dynamic conceptual units occurring in the game industry in order to create terminological descriptions that may suit the needs of both game industry professionals and final game users.

Pasanen's article is based on her doctoral thesis on methods for extracting terminological information from Finnish and Russian texts in the subject domain of maritime safety. She describes and compares the results of empirical studies of manual and semi-automatic term extraction methods, carried out by students and experts, respectively, in Finnish and Russian. In addition, the use of markers of terminological information (so-called knowledge probes) is discussed on the basis of an empirical study of Finnish and Russian corpora.

Nina Pilke,
University of Vaasa,
Finland

Margaret Rogers,
Surrey University,
UK

Birthe Toft,
University of Southern Denmark

Päivi Pasanen
University of Helsinki

EXTRACTING TERMINOLOGICAL INFORMATION FROM TEXTS

Abstract

This paper discusses the methodology of extracting information from texts from a theoretical point of view as well as from the perspective of a terminologist or a translator. The paper first deals with certain term extraction methods and an empirical study of term extraction. As the result of the study, the need to further develop term extraction methods is emphasized. The latter part of the paper discusses identification of other terminological information, e.g. definitions and concept relations. It is argued that terms and other terminological information must be extracted together and, consequently, a combined model is proposed. The paper is based on my doctoral thesis (Pasanen 2009), which discusses some methods for extracting terminological information from Finnish and Russian texts in the subject domain of maritime safety.

1 INTRODUCTION

In the modern world, the number of specialized texts is growing rapidly in every specialized domain. Due to the information boom, manual information management is becoming impossible. Therefore, all possible methods to manage information automatically or semi-automatically need to be employed. Terminology work in the Wüsterian tradition has been oriented towards standardization. The methods applied in compiling normative glossaries are well established and widely used. The development of methods for descriptive terminology work has started recently and new methods have been introduced. However, the research has mainly been English language oriented. Therefore, there are no satisfactory language-independent methods for extracting terminological information from texts.

The purpose of this paper is to discuss the methodology of extracting information from texts from a theoretical point of view and, consequently, to form the basis for a further improvement of the methods currently used for the extraction of terminologically relevant information. The basic concept in our study is the concept of terminological information. The concept terminological information can be defined as all the information connected with a concept and expressed lexically in texts of the specialized domain in question, i.e. terms and term-like designations, definitions and lexical or semantic hints referring to the concept relations. The study was based on the assumptions that, first, the terminological information necessary for descriptive terminology work is included in specialized texts and that, second, at least some of it can be extracted by using information extraction tools.

There are designations of concepts belonging to a specialized domain in the texts. In addition, there is information about characteristics and concept relations. This paper first deals with an empirical study of certain term extraction methods. These are manual term identification and semi-automatic term extraction, the latter of which was carried out by using three commercial computer programs. Section 2 of this paper presents the main results of the term extraction study. Section 3 discusses the use of certain words, phrases or other means which can be read as markers of terminological information in texts. Some researchers call these markers knowledge probes. This part of the study originated from the observation that authors use certain words or phrases to emphasize terminological information. Also, the fact that the principle of using such means is language independent encouraged us to study them further. In our study, the research material consisted of electronically readable specialized texts in the subject domain of maritime safety. A textbook, conference papers, research reports and articles from professional journals in Finnish and in Russian were included. In section 4, the results of the study are summarized.

2 TERM EXTRACTION STUDY

The aim of the term extraction study was, firstly, to determine differences in term extraction patterns between a subject group with knowledge of the specialized domain in question and a subject group without this knowledge. A group of Finnish translation students and a group of Finnish maritime students as well as a group of Russian language students and a group of Russian maritime students were invited to participate. The results were then compared with the lists of Finnish and Russian terms compiled by Finnish and Russian maritime experts and terminologists. Secondly, the aim of the study was to test the performance of three commercial term extraction tools as seen from the terminologist's point of view. A further aim was to investigate the core terms in the corpus on the one hand, and the terms which are difficult to identify on the other hand.

The term extraction study was conducted by comparing the term candidate lists produced by students and by means of term extraction tools with an established benchmark, i.e. the term lists produced by domain experts and terminologists. The term lists used as a reference list were compiled manually by three Finnish-speaking and three Russian-speaking experts. They had knowledge of the maritime field and/or in the field of terminology. In order to diminish the influence of individual preferences, a group work-method was employed to carry out the term extraction task. The task of compiling the term lists was performed in phases. During the first phase of the task, the experts extracted terms individually. During the second and third phases, the experts had to comment on the suggested term lists and to agree on the final list. Interestingly, after the first round just one third of the suggested terms were extracted with one voice by all three Finnish experts. In the preliminary Russian term list only 12 per cent of suggested terms were extracted unanimously. This study shows that to make a distinction between a term and a non-term is not easy for a domain expert or an experienced terminologist, not to mention for lay people. As a result, a mutual agreement on valid terms was reached. The final list of Finnish terms includes 189 terms and the final list of Russian terms consists of 107 terms.

In our study, the manual term extraction was conducted by Finnish and Russian maritime and language students. The students comprised four groups, each of which included 11 students. In the semi-automatic term extraction, the tools NaviTerm 2.0 by Connexor, MultiTerm Extract by SDL Trados and Masterin by Masterin Innovations were used. All of them extract terms from texts in Finnish, but just one of them, namely MultiTerm Extract "reads" texts written in Russian. In both the manual and semi-automatic term extraction methods, the research material consisted of one Finnish conference paper comprising 2,366 words (Wihuri 2002) and one Russian article consisting of 2,794 words (Pričkin & Kljuev 2000). The results of term extraction were compared and the recall and precision of the methods were evaluated. Here, recall is the number of extracted terms divided by the number of all terms in the text, i.e. the number of terms extracted by the experts. Precision is the number of extracted terms divided by the number of term candidates extracted. Before discussing the study on term extraction in more detail, we give a short introduction to the term extraction methods applied.

The manual term extraction method is based solely on human resources and no information technology is employed. The term candidate list produced by a domain expert or an experienced terminologist may well be called a term list. However, the human factor is always present in manual term extraction. Experience shows that manual term recognition is dependent on individual choices even if the test subjects have the same educational background (Pasanen 2009: 105). Besides this, the method is time consuming. Consequently, a terminologist or a translator might wish to employ term extraction software for compiling term lists in a specialized domain.

In the semi-automatic term extraction method, information technology is employed. Since the 1980s the combined efforts of engineers and linguists have produced a selection of term extraction tools. In spite of all efforts to model natural language, term extraction still remains semi-automatic, since there are no computer programs on the market which could produce acceptable term lists without any editing afterwards. The semi-automatic term extraction tools utilize algorithms which are based on either statistical or linguistic parameters of terms in texts. Statistical semi-automatic term extraction methods are based on counting the frequencies of words and word sequences. MultiTerm Extract, for example, mainly utilizes the statistical method. The advantages of statistical methods are that the principle applied is language independent and that also single-word terms are detected. The other type of semi-automatic term extraction software, linguistic term extraction software, is based on recognition of term patterns in texts. These methods are language dependent. The more advanced form of semi-automatic term extraction is the so-called hybrid method. Besides statistical counts, they use morphological and syntactic methods of text analysis and detect nominal phrases with the ideal length of 1 to 3 words. NaviTerm 2.0,

and Masterin term extraction software are examples of tools which utilize the hybrid method of term detection. The NaviTerm 2.0 software algorithm is reported in detail in Lahtinen (2000). Since they employ linguistic analysis, the NaviTerm 2.0 and Masterin term extraction tools are language dependent. At the time of the term extraction study, these tools could not extract term candidates from texts written in Russian.

Due to a combination of two basically different methods, hybrid methods have given better results than purely statistical or linguistic methods. Nevertheless, practice and research both suggest that the term candidate list produced automatically will contain noise, i.e. a lot of invalid term candidates, and silence, which means that many valid terms are missing from the list. A great variety of term extraction tools and different approaches have been thoroughly reviewed in Cabré et al. (2001). However, due to the lack of a common test bench, comparison of the tools is difficult. Vivaldi and Rodríguez (2007) joined the discussion by presenting a model for the evaluation of term extraction systems.

2.1 The experts' term list

It has been argued that the majority of terms are nominal phrases consisting of two words. However, according to our term extraction study (Pasanen 2009: 130), about half of the Finnish terms consist of one word only. This can be explained by the agglutinative nature of Finnish. This language characteristic is manifested in long one-word compound terms, e.g. merenkulkukoulutus 'maritime education'. Based on the Finnish term list compiled by one domain expert and two terminologists, more than half (61 per cent) of the Finnish terms are single-word terms, about one third of them consist of two words and just 10 per cent have more than two words. Unlike the Finnish terms, most of the Russian terms consist of two or more words. Only one fourth of Russian terms extracted by the experts consist of one word only and more than one third of the terms consist of two words, e.g. морской транспорт 'sea traffic'. Equally high is the number of terms having at least three words. By a word we mean a string preceded and followed by a space, and, unlike in some other studies, prepositions are counted as words. Furthermore, 66 per cent of the Finnish terms and 52 per cent of Russian terms occur only once in the source text. Therefore, a term does not necessarily have two or more occurrences in a text as is often assumed in semi-automatic term extraction systems (see e.g. Sewangi 2001: 94; L'Homme et al. 1996: 299, 300, 309).

The most common term patterns in the term list of the Finnish experts are single-word compound nouns (e.g. informaatiopalvelu 'information service'), single-word simple nouns (e.g. satama 'port') and phrases consisting of an abbreviation and a simple or compound noun (e.g. VTS operaattori 'VTS operator'). More than 60 per cent of the Finnish terms match these term patterns and about 97 per cent of them match the eight most common term patterns. As is typical of maritime Finnish and of the source text especially, the term list contains 11 English language terms (e.g. Long Range Tracking), 10 abbreviations of English words (e.g. SAR) and 24 terms that are combinations of an English abbreviation and a Finnish word (e.g. VTS hallinto 'VTS administration'). The English terms and abbreviations are included in the experts' term list even though most of them have Finnish equivalents in the text. This can be explained by the fact that English terms and abbreviations are widely used in Finnish maritime texts together with Finnish terms. Furthermore, variation is common in the Finnish term list, especially in long terms which designate new concepts. For example, the term laivaliikenteen ohjauspalvelu 'vessel traffic service' has three elliptic variants in the term list: laivaliikenteen ohjaus, liikenteen ohjaus and ohjauspalvelu. One reason for the variation might be that the Finnish designations have not yet been established. Surprisingly, the Finnish experts validated as terms some proper nouns, for example the abbreviation IMO, which refers to The International Maritime Organization.

With regard to the Russian terms, the number of different term patterns is higher due to a number of long terms including prepositional phrases. Still, if a term candidate matches the term pattern Ax Nx or Nx, in which the number of adjectives or nouns varies between one and four, the term candidate is most probably a term. Also some of the Russian terms have variants. For example, the terms контроль над морским транспортом 'sea traffic control' and контроль над судоходством 'vessel traffic control' designate the same concept. Elliptic terms are present in the Russian source text as well, for example, район A1 is the elliptic form of the term морской район A1 'sea area A1'. Nevertheless, variation is not as common in the Russian term list as it is in the Finnish term list. Moreover, unlike the list of Finnish terms, the Russian term list includes only three English abbreviations.

2.2 Comparison of term candidate lists with the experts' term list

The term extraction study was based on the assumption that experts identify terms relying on world and domain knowledge and the textual context. On the other hand, computer software counts word frequencies and/or recognizes linguistic patterns (e.g. Adj. + Noun). Keeping this in mind, the term candidate lists compiled by the students and extracted by the computer programs were compared with the lists of terms identified by the domain experts and terminologists. First, we simply counted the number of term candidates in the lists compiled by the students and compared the figures with the number of terms in reference lists. In this comparison all 44 term candidate lists were compared. The results show that the variation is remarkable. The shortest term candidate list includes only 29 term candidates, while the longest one includes as many as 429 candidates. For further comparison, we compiled a term candidate list for each student group consisting of term candidates which at least four members of that group had suggested. These combined term candidate lists had nearly the same number of candidates as the experts' term lists had terms. In this phase of the study we also compared term candidate lists produced by computer programs. These term candidate lists were the full-size lists meaning that we did not limit the number of candidates by using the noise ratio function available in the programs.

Having compiled the term candidate lists, we compared them with the experts' term lists focusing on the full matches, i.e. the terms which were extracted by the students and by means of the computer programs, as well as on undergeneration, which are the non-extracted terms, and on overgeneration, which are the invalid term candidates. These comparisons were made language by language and list by list.

2.2.1 The term candidates extracted by students

The Finnish maritime students identified 128 terms out of 189. The recall in this student group is 68 per cent. The terms identified by the Finnish maritime students are either terms having a high frequency in the text (e.g. *VTS* 'VTS'; *meriliikenne* 'sea traffic') or terms with a low frequency and a high degree of termhood according to the experts (e.g. *ahtaus* 'stevedoring'; lastaussuunnitelma 'cargo plan'; *VHF radio* 'VHF radio'). The precision is as high as 67 per cent since only 63 term candidates chosen by the Finnish maritime students are not terms according to the experts. The term candidate list of the Finnish translation students includes 238 term candidates. 137 of these are included also in the term list compiled by the experts. Thus, the recall is 72 per cent, which is slightly higher than in the group of maritime students. Half the undergenerated terms are the same as in the term candidate list of maritime students. The translation students had chosen 101 term candidates which are not terms. The precision of the term identification in this group is 58 per cent, which is quite low compared with the group of maritime students. The Finnish students identified most of the long terms. On the other hand, they also extracted long term candidates which are not terms.

The Russian maritime students identified 74 terms out of 107. This gives a recall of 69 per cent, which is one per cent higher than in the group of Finnish maritime students. But, unlike the Finnish maritime students, the Russian students extracted a lot of term candidates which are not terms. The number of 179 invalid term candidates gives a precision of 29 per cent only. In the group of Russian language students, the recall is 77 per cent, since the language students extracted 82 terms out of 107. The number of invalid term candidates is low. 79 invalid term candidates chosen by the language students correspond to a precision of 51 per cent, which is notably higher than in the group of maritime students. Unlike the Finns, the Russian students identified most of the short terms but had difficulties in the identification of long ones. This is partly explained by the variation in prepositional structures, which produce synonymous terms with different grammatical structures (e.g. *контроль за движением судов and контроль движения судов*).

Term candidate list	Language	Number of term candidates	Recall %	Precision %
kofi[1]	fi	238	72	58
mofi[2]	fi	191	68	67
koru[3]	ru	161	77	51
moru[4]	ru	253	69	29

[1] Finnish translation students, [2] Finnish maritime students, [3] Russian language students, [4] Russian maritime students

Table 1: Recall and precision of term candidate lists compiled by student groups

It has been argued that term identification is based on intuition, but actually a human compiler conducting the task of term extraction utilizes world knowledge, domain knowledge and language knowledge (cf. L'Homme et al. 1996: 293). The comparison of manually extracted term candidate lists gives support to Kobrin (1989: 23, 27), who claims that native speakers easily recognize terms from a text regardless of education. In fact, the subject studied has less effect on the result than individual assumptions about a term. The recall was higher in the subject groups of language students. This result suggests that term recognition is somewhat easier for language students than for the students of the specialized domain. On the other hand, the Finnish maritime students did not extract a large number of invalid term candidates. However, this is exactly what the Russian maritime students did.

In spite of the fact that specialized education is not crucial in term extraction, there are some differences in the term candidate lists which can be explained by knowledge of the specialized domain or by knowledge of terminology. The most obvious cases are the Finnish term vaarallinen aine 'dangerous good' and the Russian term бедствие 'emergency'. The language students did not recognize these terms, whereas most maritime students identified them. The Finnish translation students did not recognize the terms ohjailla 'navigate', ohjaus 'navigation', rantavaltio 'port state', turvallisuus 'safety' and näkyvyys 'visibility', for example, whereas eight or nine maritime students out of eleven did identify them. The translation students possibly concluded that these terms are general-language words. The Finnish maritime students did not identify the terms horisontaalinen integraatio 'horizontal integration' and varastointi 'warehousing', for example. Probably, these terms are considered to belong to adjacent domains. The Russian maritime students did not identify, for example, the terms район A1 'area A1' and несение вахты 'watchkeeping'. These terms, like most of the undergenerated terms, are related to the ordinary duties of deck officers. Perhaps, these terms are even too common to receive the status of a term.

The assumption was that translation students would not extract proper nouns, since they should know that terms designate general concepts. However, the abbreviations IALA, IMCO and IMO, which designate maritime organizations, are in the term candidate lists, as well as in the term list compiled by the experts. The translation students probably did not know that these abbreviations refer to organizations rather than general concepts, because they did not extract the proper nouns Elbe-joki 'the Elbe', Hampurin satama 'the Port of Hamburg' and Malaccan salmi 'the Strait of Malacca', which were in the term candidate list of Finnish maritime students. Besides proper nouns, the Finnish maritime students overgenerated adjectives (e.g. kansainvälinen 'international'). The Finnish translation students, on the other hand, overgenerated phrases which may include a term, but in addition, include extra words (e.g. nykyaikainen VTS toiminta 'the running of a modern VTS').

2.2.2 The term candidates extracted by term extraction software

The computer programs produced term candidate lists which were longer than the lists compiled by experts or students. This is natural, because a candidate list produced by a computer program inevitably has some noise, i.e. overgenerated term candidates. Normally, the precision of semi-automatic term extraction varies between 30 and 70 per cent (L'Homme et al. 1996: 303–307). Consequently, a great number of the extracted term candidates are not terms. All three term extraction tools included in the study allow refining the search. Since a high recall usually results in low precision and vice versa, the choice has to be made in favour of either recall or precision. In our study the term candidate lists were produced aiming at a maximally high recall. With the default settings NaviTerm 2.0 produced a list of 694 term candidates from the Finnish text. The full match is 129 terms and undergeneration 59 terms giving a recall of 68 per cent. The number of overgenerated term candidates is 565 which gives a precision score of 19 per cent. The term candidate list produced by Masterin has almost the same number of

candidates. Out of 690 term candidates, 116 are terms. Therefore, the recall and precision are lower, 61 and 17 per cent respectively.

In principle, MultiTerm Extract software is language independent and "reads" both Finnish and Russian texts. The term candidate lists were produced with the default settings, except that the noise ratio was set at 75 per cent. Using this setting the Finnish term candidate list includes 306 candidates and the Russian candidate list comprises 365 term candidates. As desired, these numbers are somewhat higher than the number of terms included in the reference lists. The program has a lemmatizer, but this feature is not completely developed especially as far as Finnish language is concerned and therefore there were a number of inflected forms of term candidates with the same basic form. For example, the basic form alueellinen 'regional', the genitive form alueellisen and the plural partitive form alueellisia are all included in the term candidate list. After completing the lemmatizing process manually, the Finnish term candidate list comprised 255 and the Russian term candidate list 349 lemmatized candidates. 48 of the Finnish candidates and 29 of the Russian candidates really were terms. This gives recall rates of 25 and 27 per cent. The program managed to extract most of the terms with a high frequency in the source text, for example, meriliikenteen ohjausjärjestelmä 'sea traffic management system' and контроль над судоходством 'vessel traffic control'.

If the terms occurring only once in the source texts are ignored, both the recall and precision of term extraction software are higher. In the case of NaviTerm 2.0, the recall is as high as 86 per cent and the precision is 34 per cent. For MultiTerm Extract, the recall of Finnish terms is 71 per cent, but precision remains at 19 per cent. Thus, the program developers' desire to ignore low-frequency terms is understandable.

Term candidate list	Language	Number of term candidates	Recall %	Precision %
NaviTerm 2.0	fi	694	68	19
Masterin	fi	690	61	17
MultiTerm	fi	255	25	19
MultiTerm	ru	349	27	8

Table 2: Recall and precision of term extraction software

The NaviTerm 2.0 and Masterin term extraction tools extracted 80 per cent of the terms that occur twice in the source text, and almost all the terms that have at least three occurrences in the source text. The high recall of high frequency terms suggests that most of the undergenerated terms have a low frequency. When comparing the NaviTerm 2.0 term candidate list with the Finnish term list, it can be noted that undergenerated terms are mainly one or two word terms with a low frequency in the source text. The largest group of undergenerated terms are terms that do not occur in the text in the form which they have in the reference list. For example, the simple noun hinaus 'towing' occurs in the text as a part of the compound noun hinauspalvelu 'towing service', and the compound noun luotsipalvelu 'pilot service' occurs in the elliptical form as a part of a longer phrase luotsi- ja hinauspalvelu 'pilot and towing service'. Of course, computer programs cannot detect strings which do not occur in the text. 13 of the undergenerated terms are combinations of an abbreviation and a noun, for example, VTS operaattori 'VTS operator', which are written incorrectly without a hyphen in the source text. The term extraction tool recognized the abbreviation and the noun but not the combination. Undergenerated long terms consist of four words and the connector ja 'and'. These terms are too long to match the algorithm.

The Masterin term extraction tool performed better with long terms, but undergenerated one-word terms. The biggest difference is the higher number of undergenerated low-frequency terms, i.e. terms that occur only once in the source text. Terms consisting of an abbreviation and a noun and English terms, which were difficult to recognize for NaviTerm 2.0, were not problematic for Masterin. On the other hand, the program undergenerated compound noun terms with a low frequency (f = 1). Interestingly, the term ohjaus 'management', which occurs eight times, and the term ohjausjärjestelmä 'management system', which occurs four times, are not in the term candidate list. For the NaviTerm 2.0 and Masterin term extraction tools, undergeneration is not a real problem, although about one third of valid terms were undetected.

The low precision suggests that the tools produce a large number of high-frequency words or phrases that are not terms. Actually, this is not the case when NaviTerm 2.0 and Masterin are concerned, since more than 80 per cent of overgenerated term candidates occur only once in the source text. These are mainly general-language nouns that match the common term pattern N (e.g. kehitys 'development'; määritelmä 'definition') or phrases which are either too long or composed of general language words (e.g. asiantuntijoiden yleinen mielipide 'general opinion of experts'; suora yhteys 'direct connection'). The explanation for overgeneration is the grammatical structure of these term candidates. For example, the term candidate suora yhteys has the structure adj. + noun, which is a common structure of a term. The number of overgenerated term candidates is almost equal, but the NaviTerm 2.0 tool overgenerates mostly one-word candidates, nouns or compound nouns, whereas Masterin overgenerates mostly two-word candidates. According to Lahtinen (2000: 12, 174, 182) the NaviTerm 2.0 algorithm includes an inverse document frequency (IDF) measure, the aim of which is to distinguish terms from general-language words, but the term candidate list shows that knowledge in the specialized domain is still needed.

The MultiTerm Extract term extraction tool undergenerated heavily both Finnish and Russian terms. About 75 per cent of Finnish terms and 73 per cent of Russian terms remained undetected. The tool performs well when the frequency is three or more, but does not detect terms if their frequency is lower than two. The recall was highest in the extraction of short Finnish terms and lowest in the extraction of long Finnish and Russian terms, probably due to a low frequency, which is partly caused by term variation. Since the statistical approach has been predominant in the design of the tool, it does not recognize term patterns. Consequently, most of terms that have the usual term pattern, abbreviations included, go undetected. Patterns of complex terms are missing, as could be expected on the basis of term length analysis. On the other hand, the tool extracts plenty of Finnish and Russian high-frequency term candidates that can be classified as general-language words. Many Finnish single-word term candidates are verbs, adjectives or adverbs which are not represented in the experts' list of most common term patterns (e.g. tulla 'come'; taloudellisia 'economical'; kuitenkin 'however'). From the Russian source text the tool overgenerates adjectives (e.g. технический 'technical') and single-word nouns which are general-language words rather than terms (e.g. обеспечение 'ensuring'). Many overgenerated term candidates are two-word phrases, which consist of an adjective and a noun or two nouns (e.g. точка зрения 'viewpoint'). Thus, structurally they could be terms. Similarly, the tool extracts abbreviations, but unfortunately the wrong ones. In the term candidate lists generated by MultiTerm Extract, more than 80 per cent of Finnish term candidates and over 90 per cent of Russian term candidates were overgenerated.

The NaviTerm 2.0 and MultiTerm Extract tools rank the term candidates according to their termhood. Thus, a terminologist or a translator might wish to reduce noise using an advanced search which ignores those term candidates which get the lowest scores (less than 0.1 scores in the case of NaviTerm 2.0 and less than 75 in the case of MultiTerm Extract). However, our study shows that the noise ratio remains at a high level and the silence ratio is a lot higher than with the default settings (Pasanen 2009: 120, 124). If a high recall is the aim, default settings are the best choice, even though this means a lot of manual editing afterwards.

2.3 The core terms of maritime safety

One of the assumptions in the study was that some of the terms can be called core terms, since they have a high degree of termhood and stability. Furthermore, it was assumed that both human compilers and computer programs easily identify these terms in a text. The results of the study show that there is a certain group of terms which were extracted unanimously by the experts and the majority of the students. The Finnish core terms are single-word nouns, single-word compound nouns, abbreviations or combinations thereof. The Russian core terms are single-word nouns, abbreviations or nominal phrases consisting of an adjective and a noun. Both Finnish and Russian core terms occur at least three times in the source text. Some of these terms may also belong to general language (e.g. alus 'vessel'; satama 'port'; onnettomuus 'accident') or have a low frequency of occurrence (e.g. luotsi 'pilot'), but utilizing world knowledge, domain knowledge and contextual knowledge the human compilers identified them to be terms in the particular domain. These terms are old and well established in the language and have no variants. Apart from these core terms, the experts and students recognized concepts which no doubt belong to the domain, but synonymy and variation is present (e.g. alusliikenne, laivaliikenne 'vessel traffic'). These terms gave the term candidate lists a personal touch, which is reflected in the choice of term candidates. They form a circle around the core of the domain terminology. The next circle is the

group of terms which are closer to the adjacent domains (e.g. varastointi 'warehousing'). The experts had to negotiate about the termhood of these terms.

The semi-automatic term extraction tools also performed better with the core terms than with the full term list. In the term candidate list generated by NaviTerm 2.0, the recall of core terms is high, 83 per cent, but, on average, the scores are not higher than those of other terms or term candidates. In the Finnish term candidate list produced by MultiTerm Extract, the recall of core terms is 40 per cent, which is twice as high as the recall of all terms. Here, the recall and precision were calculated using a reference list that contained those 52 terms which were extracted by at least seven students from both student groups and by all the Finnish experts during the first phase of the term extraction task.

Based on the results of our study, it might be suggested that term extraction tools of the future should include a semantic component. These kinds of tools would utilize existing ontologies and terminological information, e.g. information on concept relations, embedded in the context of a term candidate to determine whether the candidate belongs to the given specialized domain. Furthermore, to avoid starting every extraction process from scratch it is necessary that the term extraction tool memorizes domain-specific core terms identified during the previous term extraction projects. At present, terminology work may start with the semi-automatic extraction of core terms followed by extraction of terminological information using other methods. In the following section we present one of these methods, which is based on the use of so-called knowledge probes.

3 KNOWLEDGE PROBES AS INDICATORS OF TERMINOLOGICAL INFORMATION

Recently, term extraction has been incorporated in the extraction of knowledge-rich contexts semi-automatically. Knowledge-rich contexts express conceptual information, i.e. information on concept relations or characteristics. (See e.g. Meyer 2001.) Our study gives support to the notion that it is reasonable to study terms and conceptual information together, because in practical terminology work the conceptual information and the term designating the concept have to be connected. A plain term list as such, unless linked to other terminological information, does not offer all the information necessary for building a concept system. Therefore methods to identify terminological information from texts are needed. One of these methods is the use of so-called knowledge probes for the extraction of concept relations.

Knowledge probes are linguistic phrases, punctuation marks or typographic means which indicate terminological information in texts. For example, in the sentence "Windows NT is a 32-bit operating system" the knowledge probe is the combination of the verb is and the article a which indicates the generic concept relation between the subordinate concept Windows NT and the superordinate concept operating system (Kavanagh 1995: 25). Besides generic concept relations, knowledge probes may indicate synonymy, partitive relations or associative relations, for example. The principle of knowledge probes is language independent even though the actual linguistic phrases depend on languages. The biggest challenge related to the use of knowledge probes is the difficulty of formulating the search string. Furthermore, the probes vary from one domain to another and from one author to another.

Although some studies have been carried out on knowledge probes in English texts (see e.g. Grinsted 2000; Kavanagh 1995), little information is available on the functioning of them in texts written in other languages, like Finnish or Russian. We therefore decided to test this method with research material consisting of authentic, informative or didactic domain-specific texts written by domain experts in Finnish or in Russian. The material includes machine readable seminar and conference papers, research reports, a text book, action plans and articles published in a professional journal. The size of the Finnish corpus is 74,215 words. The Russian corpus is somewhat smaller, including 41,025 words. Since the aim of the study was to research the use of knowledge probes for extracting terminological information as it is used in normal communication in the domain of maritime safety, we excluded statutory texts, although abundant and easily accessed, because to some extent their language does not often occur naturally in the specialized domain and may be translated. In addition to statutory texts, dictionaries and glossaries were excluded. Although dictionaries and glossaries include terminological information in a compressed form, terms and definitions given in them are isolated from their natural textual environment in the specialized communication.

We started the study with the aim of compiling a list of Finnish and Russian knowledge probes. First, we gathered lists of possible Finnish and Russian knowledge probes manually from the corpora. After having

completed the manual search, we categorized the knowledge probes according to the type of information they indicated and produced concordances of those knowledge probes which had generated at least four valid instances from the Finnish corpus or three valid instances from the Russian corpus. By a valid instance we mean textual contexts including terminologically relevant data and by noise we mean the opposite, i.e. textual contexts which do not include terminologically relevant data. As the criteria for assessing the usability of the knowledge probes, two measures were used. These are productivity, which indicates the number of valid instances generated by the knowledge probe from the corpus, and precision, which indicates the number of valid instances divided by the number of all instances generated by the knowledge probe. Based on these measures, the knowledge probe candidates were categorized in three classes: strong, good and poor ones. The strong knowledge probes generated more than 10 instances and at least half of them were valid. The productive and precise knowledge probes were chosen for the list of Finnish and Russian knowledge probes in the domain of maritime safety. In the following, we report the main findings of the searches aiming at identification of Finnish and Russian knowledge probes in the domain of maritime safety.

3.1 Finnish and Russian knowledge probes in the domain of maritime safety

Unlike in some other studies, we did not limit the search for knowledge probes to a certain type, such as verbs, for example. Consequently, the manual search in the Finnish corpus produced strong knowledge probes indicating terms, definitions, generic relations, partitive relations and associative relations. In the Russian corpus, the search produced strong knowledge probes indicating terms, definitions, partitive relations and associative relations. Unlike what was the case in English texts, punctuation marks were not strong or good knowledge probes in the Finnish or Russian corpora. Punctuation marks are productive, but the problem is the large amount of noise they produce.

The manual search did not produce any strong Finnish or Russian knowledge probes indicating synonymy; however, the search revealed that an English equivalent of a Finnish term very often is placed in parentheses. Therefore, unlike in English, in Finnish texts, parentheses can be said to indicate equivalence rather than synonymy (cf. Pearson 1998: 174).

3.1.1 Finnish and Russian knowledge probes indicating terms or definitions

The manual search followed by a concordance search produced three strong Finnish knowledge probes indicating the designation of a concept. These are the nouns käsite 'concept' and nimitys 'name', and the abbreviation ns. 'so-called'. The noun käsite proved to be the strongest of them, since with 29 instances it is productive and with a precision of 100 per cent it does not generate noise at all (see Table 3). Besides this, the term usually follows the noun käsite in a text.

Information type	Finnish knowledge probes	Number of hits and their validity	Russian knowledge probes	Number of hits and their validity
Term	käsite 'concept'	29 100%	понятие 'concept'	33 61%
	nimi/nimitys 'name'	11 64%	термин 'term'	17 65%
	ns. 'so-called'	42 71%	концепция 'concept'	14 57%
Definition	määritelmä 'definition'	16 88%	представлять собой 'be'	17 94%
	tarkoittaa 'mean'	28 75%		

Table 3: Strong knowledge probes indicating terms or definitions in the Finnish and Russian test corpora

The search in the Russian corpus also produced three strong knowledge probes indicating the designation of a concept. These are the nouns понятие and концепция, which are the Russian equivalents for the Finnish knowledge probe käsite, and the noun термин 'term'. The most productive one of them is the noun понятие. The size of the corpora may explain the low productivity of the nouns termi and термин, since the English counterpart the term has proved to be a strong knowledge probe in English corpora (Pearson 1998: 133; Sierra & McNaught 2000: 10). In the Russian corpus, the term indicated by the knowledge probe термин may be placed in inverted commas:

Example 1:

> - - 11% не понимают термина «безопасная скорость» - - (Bezop: 7)
> [- - 11 per cent do not understand the term ”safe speed” - -]

In Finnish and Russian the search is complicated, since both Finnish and Russian knowledge probes have a number of inflected forms. In the case of the noun käsite, the search string is *käsite/*käsittee*/*käsitettä, in which the wildcard character indicates that the word käsite might be connected to a term with or without a hyphen. In the following example the hyphen is missing after the term meriturvallisuus ‘safety at sea’:

Example 2:

> Olen useimmissa meriturvallisuutta koskevissa esityksissäni jakanut meriturvallisuus käsitteen alusturvallisuuteen, väyläturvallisuuteen ja meripelastukseen. (hv98: 1)
> [In most of my presentations about safety at sea, I have divided the concept safety at sea into ship safety, fairway safety and sea rescue.]

To avoid long search strings, not all the inflected forms are included in the strings. Some cases could be left out of the search string on the basis of manual search.

Research in English corpora shows that the verb define indicates a definition (Pearson 1998: 140). Consequently, the Finnish equivalent määritellä and the noun määritelmä are obvious knowledge probes. In the Finnish test corpus the noun määritelmä is more productive with 16 hits. Since 14 of the instances are valid, it is precise, also. In the category of knowledge probes indicating definitions, the English knowledge probe means (Sierra & McNaught 2000: 10) has a Finnish counterpart, namely the verb tarkoittaa ‘mean’, which generated 28 instances. Since 21 of them are valid, the verb is classified as a strong knowledge probe. In addition, when using these knowledge probes, the search strings are easy to formulate. As the manual search indicated, the search strings may be formulated as määritelmä*, since the stem of the word remains unchanged in the necessary inflected forms, and as tarkoittaa/tarkoitetaan, where both active and passive forms are present. These knowledge probes generate incomplete definitions, which usually lack the broader concept:

Example 3:

> - - ensimmäinen lähellä oleva määritelmä saattaisi olla ”VTMIS parantaa merikuljetusten tiedonkulkua”. (pw02: 6)
> [- - the first appropriate definition might be “VTMIS improves the information flow in the field of maritime transportation”]

In the Russian corpus, the only strong knowledge probe indicating a definition is the verb представлять собой ‘be’, which is precise, and the concept to be defined either precedes the verb or follows it. Besides this, most of the definitions are full generic definitions. The English knowledge probe called (Sierra & McNaught 2000: 10) has a Russian counterpart, the verb называться. With seven valid instances out of ten altogether, this knowledge probe is precise but not exceptionally productive. In the Finnish corpus, there was no counterpart for this verb. In addition, there is one Russian knowledge probe which doesn’t have an equivalent in Finnish or English. It is the combination of a dash and the pronoun это ’this’. It can be translated as the verb to be. This knowledge probe produces full generic definitions:

Example 4:

> Танкер – это судно, предназначенное для перевозки жидких грузов, имеющих свободную поверхность. (Bezop: 98)
> [A tanker is a vessel designed for transportation of liquid cargoes with a free surface.]

3.1.2 Finnish and Russian knowledge probes indicating generic or partitive relations

The manual search in the Finnish corpus generated nine knowledge probe candidates indicating a generic relation, but, surprisingly, only one of these is classified as a strong one, because most of the instances are invalid for the purpose of extracting generic relations. Thanks to its high precision the pronoun muu,

'other' – in plural phrasal form ja muut 'and other' – is the strongest knowledge probe in this category (see Table 4). Due to inflection the search string was formulated as ja muut/ja muiden/ja muita/myös muita.

Information type	Finnish knowledge probes	Number of hits and their validity	Russian knowledge probes	Number of hits and their validity
Generic relation	ja muut 'and other'	29 55%		
Partitive relation	jakaa 'divide'	20 60%	состоять из 'consist of'	17 82%
	käsittää 'contain', 'comprise'	19 58%		
	koostua 'consist of'	15 73%		

Table 4: Strong knowledge probes indicating generic or partitive relations in the Finnish and Russian test corpora

Also, the subordinating conjunction kuten 'such as' is productive and, consequently, a useful knowledge probe. It also keeps company with terms, but more than half of the instances are invalid. Like its English counterpart such as, kuten is preceded by a superordinate concept and followed by a list of subordinate concepts (cf. Grinsted 2000: 40–42):

Example 5:

> Itse liikenteen keskeisimmät vaikutusmuodot ovat emissiopäästöt ja erilaiset päästöt mereen, kuten harmaa vesi, musta vesi, pilssivesi ja painolastivesi. (smr02: 55)
> [The most central effects of traffic itself are emissions and different kinds of discharges to the sea, such as greywater, blackwater, bilge water and ballast water.]

The search in the Russian corpus gave similar results. The search produced a list of 12 Russian knowledge probes indicating generic relations. None of them is classified as a strong one. Most of them are productive, but have low precision. The strongest ones are the noun вид 'kind' and the adjective другой 'other' in its plural form другие, which is the Russian counterpart of the English knowledge probe and other indicating a generic relation (Grinsted 2000: 40–42).

The English knowledge probe include (Bowker & Pearson 2002: 219; Grinsted 2000: 40–42; Kavanagh 1995: 25, 27) has two Finnish counterparts, the verbs sisältyä and sisältää, but these verbs indicate a partitive rather than a generic relation. On the other hand, the English knowledge probes comprise, describe, consist(s) of, is/are defined as, is/are called/known as, denote(s) and designate(s) (Grinsted 2000: 40–42; Pearson 1996: 818–823) do not have Finnish counterparts. The English prepositional phrases a kind of and a type of function as knowledge probes in English corpora (e.g. Bowker & Pearson 2002: 219). The Russian counterpart вид is productive but generates more invalid than valid instances. The Finnish counterpart tyyppi and its Russian equivalent тип are in the lists of knowledge probes, but they are not outstandingly productive or precise. The English knowledge probes e.g. and for example (Grinsted 2000: 40–42; Pearson 1998: 124) have a Finnish counterpart kuten esimerkiksi, which is in the list of knowledge probes, but due to low productivity it is not classified as a strong one.

On the basis of research in English corpora, it is reasonable to suggest that the Finnish verb olla 'be' in its present tense singular on 'is' or plural form ovat 'are' often indicates a concept relation between a superordinate concept and subordinate concepts. However, the verb olla is somewhat problematic, since it is one of the most frequently used words in the Finnish language (Saukkonen et al. 1979: 41). Therefore, the verb does not always function as a knowledge probe. Consequently, a lot of unnecessary information or "noise" is extracted by using this knowledge probe (cf. Pasanen 2010: 152). In the Russian corpus the possible equivalent, the existential verb есть, 'be', does not function as a knowledge probe, because it is rarely used in texts.

Research in English corpora implies that knowledge probes are applicable especially to the search for generic relations. Somewhat surprisingly, the Finnish corpus was productive in knowledge probes indicating partitive relations. The corpus generated three strong knowledge probes indicating this relation type. These are all verbs. The most productive of them are the verbs jakaa, 'divide', and käsittää 'contain' or 'comprise', which generated 20 and 19 valid instances with the search string voidaan jakaa/jaettu and käsittä*. More than half of the instances are valid. The search string koostuu formulated

from the verb koostua 'consist of' generated only 15 hits but 73 per cent of them are valid. The Russian equivalent состоять из (with the search string состо* из) is the only strong knowledge probe in the Russian corpus. It is not especially productive but has high precision. Unlike the English equivalent consist of, which may indicate partitive or generic relation (see Pearson 1998: 146–147), the Finnish knowledge probe koostua and the Russian counterpart состоять из indicate a distinctively partitive relation. Unfortunately, some of these knowledge probes, for example, koostua and käsittää, seem to be dependent on the author. Luckily, these Finnish and Russian verbs keep company with terms. Usually, they are preceded by the term designating the broader concept and followed by the term designating the narrower concept, as can be noted in examples 6 and 7:

Example 6:

> Liikenteenjakojärjestelmä koostuu liikenteenjakovyöhykkeistä, joiden avulla vastakkaisiin suuntiin kulkevat laivat ohjataan omille liikennekaistoilleen. (smr02: 70)
> [A traffic separation scheme consists of traffic separation zones, which are used to direct vessels sailing in opposite directions to the correct traffic-lanes.]

Example 7:

> Осушительная система состоит из насосов, трубопроводов, клапанов, приемных сеток, устройств сигнализации и замера уровня воды. (Bezop: 49)
> [The drying system consists of pumps, pipelines, valves, sieves, as well as apparatus for warning about and measuring the water level.]

The English noun part is regarded to be a strong knowledge probe which indicates the partitive relation (Grinsted 2000: 43; Kavanagh 1995: 27). This result is at variance with the results of our study, since the Finnish equivalent osa and the Russian equivalent часть produced only a small number of valid instances.

3.1.3 Finnish and Russian knowledge probes indicating associative relations

In the domain of maritime safety the knowledge probes proved to be especially productive when searching for associative concept relations, especially causal relations. This holds for both Finnish and Russian texts. Causal relations thus seem to be a domain-related characteristic of the texts. The manual search in the Finnish corpus followed by a concordance search produced a list of nine strong Finnish knowledge probes indicating causal relations (see Table 5).

Information type	Finnish knowledge probes	Number of hits and their validity	Russian knowledge probes	Number of hits and their validity
causal relation	aiheuttaa 'cause'	118 83%	привести 'cause'	22 77%
	aiheutua 'result from'	48 69%	приводить 'cause'	15 80%
	johtaa 'result in'	32 69%	вызвать/вызывать 'cause'	14 93%
	pyrkiä 'aim at'	22 68%		
	syy 'cause'	38 58%	в результате 'result'	42 62%
	seuraus 'consequence'	32 53%	из-за 'due to'	13 62%
	tarkoitus 'intention'	36 56%		
	tavoite 'goal'	41 54%		
	tehtävä 'function', 'purpose'	13 100%		

Table 5: Strong knowledge probes indicating causal relations in the Finnish and Russian test corpora

By far the strongest probe is the verb aiheuttaa, 'cause', which is both productive and precise. Moreover, the verb is preceded by the cause concept and followed by the consequence concept, as in example 8:

Example 8:

> Suomen olosuhteisiin soveltuisi ns. Tuck & Taylorin menetelmä matalikon aiheuttaman imun (squat) laskemiseksi. (kiv97: 82)

[In the Finnish circumstances, the so-called Tuck & Taylor method for calculating squat would be suitable.]

The possible Russian equivalents, the aspectual verb pairs приводить, привести and вызывать, вызвать are precise as well, i.e. almost all the instances include information about causal relations between concepts. These results are not surprising, since the English verb cause has proved to be a strong knowledge probe (Bowker & Pearson 2002: 219; Grinsted 2000: 45–47). Also the Finnish verbs aiheutua 'result from' or 'be caused by', johtaa 'result in' or 'lead to' and pyrkiä 'aim at' are strong knowledge probes, but the verb aiheuttaa alone produces more valid instances than all the aforementioned verbs together. The verbs aiheutua and johtaa have English counterparts, namely the verbs result from and result in (Bowker & Pearson 2002: 219; Grinsted 2000: 45–47). The Finnish verb johtaa is a strong knowledge probe, even though it is polysemous having, for example, the equivalents conduct, derive, manage and transmit. The English knowledge probe produce (Bowker & Pearson 2002: 219) has a Finnish equivalent tuottaa, which is a knowledge probe candidate, but in our corpus it produced mainly noise.

Besides productive verbs, some nouns are also classified as strong knowledge probes indicating causal relations. These are the nouns syy 'cause', seuraus 'consequence', tarkoitus 'intention', tavoite 'goal' and tehtävä 'function' or 'purpose'. The disadvantage of using these nouns as knowledge probes is that the search string is much longer than when using verbs, for example, syy/syyt/syyn/syynä/syyksi/syitä/syyhyn.

In the Russian corpus, the search produced five strong knowledge probes indicating causal relations. These are the verbs приводить, привести, вызывать and вызвать mentioned before and the noun результат 'result' which is productive and precise in the phrasal form в результате:

Example 9:

Известно, что подавляющее большинство пробоин происходит в результате столкновений или посадки на мель. (Bezop: 99)
[It is a well-known fact that the great majority of leakages result from collisions or grounding.]

In the Finnish and Russian corpora, there may often be more than one cause and consequence in the relation. Causes and consequences may also be alternative, as in example 9 above. Interestingly, in the list of Russian knowledge probes indicating causal relations, one preposition is classified as a strong knowledge probe. The preposition из-за 'due to' produces 13 instances, 8 of which are valid.

Causal relations are related to activities and phenomena involved in these activities. It's therefore natural that the knowledge probes indicating this type of relations are verbs. Equally naturally, the influencing factor is connected to activities by verbs, namely the verbs vaikuttaa 'influence' or 'affect', parantaa 'improve' and vähentää 'reduce', which are productive and precise, as can be seen in Table 6 below.

Information type	Finnish knowledge probes	Number of hits and their validity	Russian knowledge probes	Number of hits and their validity
influencing factor	vaikuttaa 'influence' parantaa 'improve' vähentää 'reduce'	85 62% 29 55% 52 52%	влияние 'effect' за счет 'because of'	31 65% 14 79%
activity-method relations	käyttää 'used for'	13 100%		
instrumental relations	avulla 'by means of' keino 'means'	120 53% 26 73%	применять 'apply' использовать 'use' служить 'function as' предназначить 'intend', 'design' метод 'method' путем 'by means of'	17 71% 11 73% 13 69% 10 70% 11 73% 16 67%

Table 6: Strong knowledge probes indicating activity relations in the Finnish and Russian test corpora

In the Russian corpus, however, the strongest knowledge probes indicating an influencing factor were not verbs. This can be explained by the typical characteristic of Russian language to express meanings with a combination of a "colourless" verb and a noun instead of an expressive verb, which is the more common way in Finnish. So, instead of using the verb влиять, the author of the following sentence has used the V + N combination оказывать влияние 'have an effect':

Example 10:

> Груз как элемент транспортной системы оказывает большое влияние на безопасность плавания судна. (Bezop: 17)
> [Cargo as an element of a transportation system has a significant effect on the safe navigation of a vessel.]

The noun влияние 'effect' was more productive than the phrasal expression за счет 'because of', 'owing to' or 'due to', which is another strong Russian knowledge probe indicating an influencing factor.

Research in English corpora shows that the verbs used for, used to, employed to and designed to are employed to indicate instrumental relation (Bowker & Pearson 2002: 219; Grinsted 2000: 45–47). A search in the Finnish corpus, however, did not produce any verbs indicating instrumental relation, since the verb käyttää 'use' proved to be a strong knowledge probe indicating the relation between an activity and a method. But the study did produce two strong nouns for searching information about activity relations, namely the nouns keino and apu, both of which can be translated as means in English. The more productive of them is apu, when the search string is formulated as avulla 'by means of'. The instrument referred to almost always precedes the knowledge probe, as in example 11:

Example 11:

> Suomen ympäristökeskus pyrkii estämään alusten tahallisia öljypäästöjä valvonnan avulla. (smr02: 71)
> [The Finnish Environment Institute endeavours to prevent intentional oil spillages by means of surveillance.]

In contrast, in the Russian corpus the verbs применять 'apply' or 'use', использовать 'use', служить 'function as', 'serve as' or 'serve for' and предназначить 'intend' or 'design' indicate a distinctively instrumental relation. These verbs are often accompanied by the preposition для – indicating the activity which the instrument is used for. The activity concept follows the preposition immediately if the preposition is connected to the verb without any "extra" words in between, as in example 12:

Example 12:

> Нижние коуши и кренгельс служат для крепления подкильных концов. (Bezop: 52)
> [Lower eyelets and grommets serve for attaching the under-keel ends.]

Besides these verbs, the nouns метод 'method' and путь in its inflected form путем 'by means of' are strong knowledge probes indicating activity relations. The latter functions well even without a term.

3.2 Evaluation of knowledge probes as a method of extracting terminological information

The aim of the research on knowledge probes is to identify knowledge-rich areas of text (Bowker 1996: 35). As mentioned above, one of the aims of the study reported here was to compile a list of Finnish and Russian knowledge probes, which could be utilized for the identification of knowledge-rich areas of Finnish and Russian texts in any specialized domain. The results of the study show that punctuation marks are the most productive knowledge probe candidates, but their precision is low. Therefore, they are not practicable in information extraction, unless the search can be delimited by adding a term into the search string. Lexical knowledge probes, instead, are precise enough to be employed in information extraction. They include a number of verbs or nouns which generate solely valid instances, i.e. instances which include terminologically valid information. Normally, they are not productive; therefore, a large corpus is needed for an effective use of these knowledge probes. Earlier research on knowledge probes

has mainly focused on verbs (see e.g. Christensen 2000). The results of our study showed, indeed, that verbs are most useful as knowledge probes. First, there is a great selection of them, second, they are productive, and third, they are fairly precise. Regardless of the language, verbs functioning as knowledge probes usually appear in the present tense and are preceded or followed by a term. Besides verbs, also some nouns, abbreviations and conjunctions deserve to be placed in the list of strong knowledge probes.

In spite of encouraging results, the use of knowledge probes for extracting terminological information involves a number of challenges. First, the search string is difficult to formulate so that it is neither too broad nor too narrow. If it is too broad, a lot of noise will be extracted. If it is too narrow, a lot of valid instances remain undetected. Christensen suggests refining the search with prepositions. In Russian this refining technique can be applied, but the so-called free word order decreases the effect of this means since a preposition is not always directly connected to the verb. Also, if the valency of a verb requires a prepositional object as an argument, the verb keeps company with the same preposition in any instance, whether it is terminologically-interesting or not. Second, the case endings and personal suffixes make the definition of the search string challenging both in Finnish and in Russian. The use of wild card results in long concordance lists containing a lot of invalid instances. The most promising solution is to add a term to the search string. However, some of the knowledge probes function quite well without a term being included in the search string, for example the verbs tarkoittaa 'mean', koostua 'consist of', aiheuttaa 'cause', johtua 'result from', käyttää 'use', представлять собой 'be', состоять из 'consist of', включать 'include', применять 'apply' or 'use', вызвать 'cause', the nouns käsite 'concept', aiheuttaja 'cause', tehtävä 'function' or 'purpose', keino 'means', компонент 'component', помощь 'means' and the abbreviation ns. 'so-called' are precise even without the company of a term. Third, one knowledge probe may function in multiple roles indicating different kinds of terminological information. For example, parentheses are a general knowledge probe, which can indicate synonymy, equivalents and definitions. According to Grinsted (2000: 47), the English knowledge probe describe* as may indicate associative or generic relations. Similarly, the Finnish verbs sisältää 'include', sisältyä 'include' and jakaa 'divide' indicate both generic and partitive relations. Fourth, the authors sometimes refer to concepts using elliptical designations. These may be interpreted wrongly to designate superordinate concepts if taken out of the textual context.

In the domain of maritime safety, Finnish and Russian knowledge probes are especially useful if the aim is to extract terminological information about causal relations. This can be explained by the fact that seafaring is closely related to technology and natural sciences, which are characterized by causal concept relations (see e.g. Meyer 2001: 296). The Finnish language differs from the Russian language in many ways; however, the Finnish knowledge probes studied here have a lot in common with the corresponding Russian knowledge probes. The explanation might be that knowledge probes are field specific. Another possible reason is that the presuppositions about the knowledge probes have directed the search towards similar words or phrases.

4 CONCLUSIONS

The results of the term extraction study suggest that specialized knowledge is an advantage in manual term identification. However, in the term candidate lists the variation between student groups was not as remarkable as it was between individual students. In each student group the recall was about 70 per cent, but the precision varied between 30 and 67 per cent. The recall was higher in the groups of Finnish translation students and Russian language students than in the groups of maritime students. This implies that knowledge of terminology methods is more useful in term extraction than specialized knowledge.

The term extraction software tested here produces term candidate lists which can be useful, but only after some editing, which has to be done manually. At best, the computer programs produce a term candidate list that has a high recall, but at the expense of precision. Furthermore, based on the study of the length, frequency of occurrence and lexical patterns of the term candidates, some improvements may be suggested to get a higher recall and precision, although undergeneration of low-frequency terms and overgeneration of high-frequency general language words and phrases still remain the biggest challenges for automated term extraction. These problems are hard to overcome with present semi-automatic term extraction methods.

The design of term extraction software is based on the assumption that a term occurs in the source text at least twice. Although the number of low-frequency terms might depend on text length, the term extraction study gives support to the claims that this assumption should be reconsidered, since almost 66

per cent of terms chosen by the Finnish experts and 52 per cent of terms chosen by the Russian experts occurred only once in the source text.

Our study gives support to earlier research showing that synonymy, term variation and ellipsis are common phenomena in languages for specific purposes. Computer programs have difficulties in recognizing these phenomena, because termhood does not depend on the form of the term as much as it depends on the use of the term in the source text. Nevertheless, the analyses indicate that there are a certain number of terms which were extracted both by the student groups and by the software. These terms we call core terms.

As the result of the study on linguistic expressions which signal concept relations, it can be noted that specialists in the field use certain expressions which indicate concept relations. In the study, a proposal for Finnish and Russian knowledge probes in the field of maritime safety was made. In the semi-automatic extraction of concept relations, the knowledge probes provide the best results if a term identified earlier is included in the search string, since knowledge probes usually occur in the vicinity of terms.

The method of extracting terminological information suggested here starts with semi-automatic extraction of core terms. For this task, any commercial term extraction software can be used. In this connection, manual editing of the term candidate list would be easier if the software had a concordance function integrated with term extraction. Alternatively, core terms can be extracted using the word list and concordance functions of a corpus program and knowledge probes. Here, core terms include foreign (English) equivalents and abbreviations. On the grounds of the term candidate list, the compiler can decide whether the source text is suitable for further information extraction. The next phase consists of extraction of definitions and concept relations using knowledge probes and combinations of core terms and knowledge probes as search strings for a corpus search. In addition to core terms, also known domain-specific and general scientific terms may be employed. In the domain of maritime safety these include, for example, the terms meriliikenne ‘sea traffic’, laivaliikenne ‘vessel traffic’ and järjestelmä ‘system’. Eventually, by repeating the searches and analyses of data generated, the preliminary conceptual structure can be outlined.

REFERENCES

Bezop = SKOPKOV, V., G. KONOPEL'KO & V. VASIL'EVA (1994). Bezopasnost' moreplavanija. Moskva: Transport.

hv98 = VALKONEN, H. (1998). Meriturvallisuus. A seminar paper presented at the Sea Safety Seminar in Helsinki, February 5, 1998.

kiv97 = Komentosiltatyön inhimilliset virheet. 1997. Helsinki: The Finnish Maritime Agency.

pw02 = WIHURI, P. (2002). Meriliikenteen ohjausjärjestelmät VTS ja VTMIS. An unpublished conference paper presented in Uusikaupunki, Finland, September 23, 2002.

smr02 = HÄNNINEN, S., R. JALONEN, T. NYMAN, A. PALONEN, K. RISKA, J. RYTKÖNEN & S. SONNINEN (2002). Suomenlahden meriliikenteen riskitekijät. A feasibility study. Espoo: School of Science and Technology.

BOWKER, L. (1996). Towards a Corpus-Based Approach to Terminography. Terminology 3(1), 27–52.

BOWKER, L. & J. PEARSON (2002). Working with Specialized Language. A Practical Guide to Using Corpora. London and New York: Routledge.

CABRÉ CASTELLVÍ, M. T., R. ESTOPÀ BAGOT & J. VIVALDI PALATRESI (2001). Automatic Term Detection. A Review of Current Systems. In Bourigault, D., C. Jacquemin, M.-C. L'Homme (eds.) Recent Advances in Computational Terminology. Philadelphia: Benjamins, 53–87.

CHRISTENSEN, L. W. (2000). Danske verber som knowledge probes i terminologisk korpusarpejde. In Nuopponen, A., B. Toft, J. Myking (eds.) I terminologins tjänst. Festskrift för Heribert Picht på 60-årsdagen. Proceedings of the University of Vaasa, Reports 59. Vaasa: University of Vaasa, 243–279.

GRINSTED, A. (2000). 'Knowledge probes' og eksempler – på jagt efter definitioner og begrebsrelationer i et korpus inden for området 'entrepreneurship'. In: Nuopponen, A., B. Toft, J. Myking (eds.): I terminologins tjänst. Festskrift för Heribert Picht på 60-årsdagen. Proceedings of the University of Vaasa, Reports 59. Vaasa: University of Vaasa, 36–51.

KAVANAGH, J. (1995). Text Analyzer: A Tool for Extracting Knowledge from Texts. Master of Computer Science Thesis, University of Ottawa.

KOBRIN, R. (1989). Lingvistièeskoe opisanie terminologii kak baza konceptual'nogo modelirovanija v informacionnyh sistemah. Avtoreferat dis. doktora filol. nauk. Leningrad: LGU.

L'HOMME, M.-C., L. BENALI, C. BERTRAND & P. LAUDUIQUE (1996). Definition of an Evaluation Grid for Term-Extraction Software. Terminology 3(2), 291–312.

LAHTINEN, T. (2000). Automatic Indexing: An Approach Using an Index Term Corpus and Combining Linguistic and Statistical Methods. Department of General Linguistics Publications No. 34. Helsinki: University of Helsinki.

MEYER, I. (2001). Extracting Knowledge-rich Contexts for Terminography. A Conceptual and Methodological Framework. In Bourigault, D., C. Jacquemin, M.-C. L'Homme (eds.) Recent Advances in Computational Terminology. Philadelphia: Benjamins, 279–302.

PASANEN, P. (2009). Merenkulun turvallisuuden koetinkiviä. Terminologisen tiedon poiminta teksteistä. Helsinki University Translation Studies Monographs 5. Helsinki: University of Helsinki. http://ethesis.helsinki.fi

PASANEN, P. (2010). Application of Terminological Methods in a Study of Maritime Safety Concepts. In Nuopponen, A., N. Pilke (eds.) Ordning och reda. Terminologilära i teori och praktik. Stockholm: Norstedts, 149–157.

PEARSON, J. (1996). The Expression of Definitions in Specialised Texts: A Corpus-Based Analysis. In Gellerstam, M., J. Järborg, S.-G. Malmgren, K. Norén, L. Rogström, C. Röjder Papmehl (eds.) Euralex '96. Proceedings I-II Papers Submitted to the Seventh EURALEX International Congress on Lexicography in Göteborg. Part II. Göteborg: Göteborg University, 817–824.

PEARSON, J. (1998). Terms in Context. Amsterdam/Philadelphia: Benjamins.

PRIČKIN, O. & KLJUEV, V. (2000). Edinaja sistema kontrolja nad sudohodstvom. Morskoj flot 5–6, 15–19.

SAUKKONEN, P., M. HAIPUS, A. NIEMIKORPI & H. SULKALA (1979). Suomen kielen taajuussanasto. Porvoo, Helsinki and Juva: WSOY.

SEWANGI, S. S. (2001). Computer-Assisted Extraction of Terms in Specific Domains: The Case of Swahili. Institute for Asian and African Studies, Publications 1. Helsinki: University of Helsinki.

SIERRA, G. & J. McNAUGHT (2000). Design of an Onomasiological Search System: A Concept-Oriented Tool for Terminology. Terminology 6(1), 1–34.

VIVALDI, J. & H. RODRÍGUEZ (2007). Evaluation of Terms and Term Extraction Systems: A Practical Approach. Terminology 13(2), 225–248.

WIHURI, P. (2002). Meriliikenteen ohjausjärjestelmät VTS ja VTMIS. An unpublished conference paper presented in Uusikaupunki, Finland, September 23, 2002.

Gianna Tarquini
PhD Programme in English for Specific Purposes
University of Naples Federico II

NEW MEDIA, NEW TERMINOLOGY CHALLENGES: THE SEMIOSIS OF ELECTRONIC ENTERTAINMENT

Abstract

Starting from the observation that modern mass culture as well as specialised communication is becoming increasingly multimodal due to the dramatic pervasiveness of new media, this paper sets out to explore conceptual representation levels used in entertainment software, a digital genre which originated in conjunction with the technological turn. While a number of new media, like web sites, have been explored from various angles, video games represent an emerging sui generis medium which has only recently gained academic attention. First, we will single out the semiotic dimensions involved in entertainment software. Not unlike television broadcasts or DVDs, video games combine the acoustic and the visual channels to create meaning. Moreover, they share iconic, textual and interaction techniques typical of web sites and software applications, combining such multimodal elements as graphics, sound, interface, gameplay and story. Next, taking a number of onscreen text instances as our object of enquiry, we will single out a number of unique extra-linguistic forms of conceptual representation relevant to video game semiosis, acknowledging their implications for specialised communication. In a changing digital landscape, the main research question is how non-verbal and dynamic conceptual units should be treated in a terminological description serving both the game industry professionals (termbases) and the final users (Help files, user guides).

INTRODUCTION

> ... Les uns pensent que l'image est un système très rudimentaire par rapport à la langue, et les autres que la signification ne peut épuiser la richesse ineffable de l'image. Or, même et surtout si l'image est d'une certaine façon limite du sens, c'est à une véritable ontologie de la signification qu'elle permet de revenir. Comment le sens vient-il à l'image ? Où le sens finit-il ? Et s'il finit, qu'y a-t-il au-delà ?
>
> (Barthes, 1964: 40, author's emphasis)

Recent studies increasingly converge in emphasizing the modern sensorial and cognitive turn towards visual modes of communication, pointing to:

> "a broad move from the dominance of writing and the written word to a relatively recent dominance of the image. That is, from a dominance on the part of the book as our primary meaning-making technology to a dominance on the part of images on screen that are colorful, that have animation, texture and dimensionality as governing technologies" (Rowsell, 2006).

This trend largely affects specialised communication, in a variety of multimodal digital forms: software applications, web sites, e-learning platforms, CD-ROMS and the like are often concerned with technical concepts either relating specifically to the field of ICT (Information and Communication Technology) or to a wide range of specialised fields. This rapidly evolving scenario addresses new issues regarding the systematisation of verbal and non-verbal terminological units with changing communicative methods and changing textual forms. Indeed, the integration of a semiotic model with terminology work and science is not alien to Wüster's communicative system (1979) and has been at the heart of epistemic discussion in the field (Budin, 1996, Järvi, 1997 and Myking, 1997). This contribution continues this discussion and analyses in particular the semiosis of electronic entertainment as an instance of multimodal specialised language, in order to discuss consequent terminological challenges.

In the modern digital era, entertainment software emerges as one of the most popular media on a global scale1, thanks to the penetration of hardware devices into our daily life, to the implementation of new

technologies and also to the rise of cyber-society and computer mediated communication (CMC). As a consequence, entertainment software has attracted growing academic attention consistent with its unquestionable cultural impact, and has been made the scientific object of study of such disciplines as narratology, ludology, media studies, semiotics, sociology, anthropology, gender studies, graphic design and programming (Salen and Zimmerman, 2006: XIX), which have provided theories and methods for the creation of a new disciplinary field, called Game Studies (ibid.). So far, however, there has been little discussion of specific linguistic and terminological features of entertainment software, partly due to some reluctance "to tackle this highly dynamic and textually complex, if not evasive, media genre" (Ensslin, 2010: 207). The written code is in fact only one component across a multimodal system including sound, dialogues, images and icons, where each semiotic sub-system employs its own language and grammar, and where, moreover, the final meaning results from the combination and not the sum of the different modes. Drawing on media studies, multimodal theory and semiotics, we will first focus on the representation levels involved in video game semiosis, as a meaningful instance of interplay between multilayered representational systems. Secondly, and as a consequence, we will discuss relevant terminological implications.

Like utility software, entertainment software involves advanced technological features, hardware/software components, and technical instructions, and is therefore concerned with (multimodal) specialised communication. As such, it fits the definition of:

> "a system for transmitting and exchanging information that employs various codes at the same time, of which human language is undoubtedly the most important, but not the only one. Other systems that are three-dimensional, two dimensional, iconic or symbolic share with human language the function of means of communication in technical and scientific contexts" (Kokourec, 1982).

Pioneering work on LSP has focused on the specialised features of ICT both in textual/structural terms (Shortis, 2001), and morphological patterns (Gotti, 1991; 2003). In practical terms, terminography serves the corporate needs of digital publishers and localisation vendors who resort to databases, terminology management applications and global content management systems to standardise the treatment of technical components (Järvi, 1997; Sandrini, 1997). Similarly, video games are subject to strict standards of terminology compliance, because hardware manufacturers and publishers have enacted strict policies to protect proprietary features, including spelling, special characters and symbols. These premises will bring to the fore the relevance of icons, visuals and other audiovisual elements in ICT, in the particular case of entertainment software, addressing the terms in which multimodal technical concepts can be represented in a way to comply with a terminological methodology. In particular, terminology records traditionally designed for the verbal system will need to integrate and classify nonverbal designations as well as dynamic hypertextual context. Such a terminological systematisation could find an application both in the creation of help files and in the scientific-professional description of entertainment software.

1 FROM PAC-MAN[2] TO EDUTAINMENT

In coming to a terminological investigation of the object of enquiry, we find the first hurdle, which is chiefly of a definitional nature. There are in fact multiple designations for the concept of entertainment software, the main occurrences being "electronic game", "computer game", "video game" (also in its orthographic variant "videogame"), or "console game"[3].

This plurality of terms reflects the array of game forms, genres, pursuits and sensorial experiences that have come under the umbrella of "video game" in a diachronic perspective of technological evolution. From Pac-Man and mobile games to flight simulation games used in professional training, edutainment and 3D movement detecting consoles, like Nintendo Wii, the game experience involves a variety of possible contents, technologies and also different age groups. In this perspective, the array of designations and possible definitions accounts for the logical difficulty in freezing the concept of an object that has undergone huge technical modifications from the 60s, when it first appeared. Despite diachronic, genre and functionality variations, the most comprehensive definition of the concept of video game has been formulated by the ludologist Gonzalo Frasca, as:

> "any form of computer-based entertainment software, either textual or image-based, using any electronic platform such as personal computers or consoles and involving one or multiple players in a physical or networked environment" (2001: 4, my italics).

This definition focuses on the essential ontological features of video games, i.e. digital implementation, interactivity and entertainment, which single them out from any other medium. For the purposes of this discussion, we will therefore opt for the terms "video game" and "entertainment software", and sometimes for the hyperonym "game", insofar as they appear to be the most unambiguous and widely accepted designations. The denomination of the object of enquiry, however, is only a premise before investigating entertainment software in a semiotic and terminological perspective.

2 THE MODES OF ENTERTAINMENT SOFTWARE

At this juncture, it is worth drawing attention to the fact that the terminological units featuring in the entertainment software communicative system are part of a complex and multilayered representational system. Far from printed texts, entertainment software encompasses verbal and nonverbal units that can be instantiated simultaneously at the level of visuals, icons, sounds and interaction (for a semiotic model of graphical user interfaces, see Järvi, 1997: 67). It is therefore necessary to grasp what is happening and at which multimodal levels before operationalising and defining technical concepts.

One of the basic tenets of the theory of multimodality is the principle that the making of meaning can be enacted through various representational systems, or modes of communication besides natural language (written or spoken). More specifically, multimodality "refers to communicative artefacts and processes which combine various sign systems (modes) and whose production and reception calls upon the communicators to semantically and formally interrelate all sign repertoires present" (Stöckl, 2004: 9). Face-to-face communication, for instance, involves the oral mode of natural language as well as nonverbal means, like gestures and body language. The print medium tends to incorporate images along with the written mode, whereas modern audiovisual media tend to "multiply the semiotic potentials by integrating moving images, language (spoken and written), sound and music" (Stöckl, 2004: 10).

Media scholars have highlighted the fact that this evolutionary trend tends to a natural process of media integration and convergence, since "no medium has its meaning or existence alone, but only in constant interplay with other media" (McLuhan, 1964: 26). Video games bring together the typical features of audiovisual media, like television and cinema, and combine them with digital interactive communication, producing the following channel-mode combinations:

Channel[4]	**Mode**[5]
Visual (Graphics)	Visuals (full motion video animation + static images) Icons Written language
Acoustic	Spoken language Sound effects (SFX) Music
Haptic (touch-related)	Formal languages/digital coding systems

Table 1: Channels and modes involved in the representational system of entertainment software

The semiotic potential of video games may involve any of these distinct representational systems. The first observation is that the totality of the visual dimensions is digitised - we may therefore refer to a "graphic channel" - unlike the visual representations traditionally featuring in photography, television or cinema. Thus, full motion pictures are actually 3D or 2D animations, photographs become screenshots, static images are a product of graphic design, and written language is displayed in a colourful and dynamic set of pixels. The purely visual modes tend to establish a mimetic relationship to reality, their direct referent being in most cases a virtual rather than a real, immanent one. The extent to which the game cyberspace imitates real spaces, or to which the representational system emulates real objects, then, is a matter of game genres and developers' creativity (Newman, 2005: 109-111). Besides real or

"hyperreal" perception (Baudrillard, 1981; Eco, 1983), each video game refers concretely to the actual console space, including hardware, peripherals, buttons and actions to be performed.

The acoustic modes are similarly digitised, in line with the current practices deployed in a number of other audiovisual media. The apparent reason for this is that video games are to all extents and purposes a piece of software which has been pre-designed and pre-programmed to be implemented on an electronic platform, and every semiotic dimension must therefore be incorporated within the source code and compiled into a build.

The haptic channel, the touch-related channel introduced by digital media (Ensslin, 2010: 106), is hard to frame within a human communication model, first because tactile communication in itself is not traditionally codified as a representational system. Second, it appears that the tactile message is only an input which can pass through other channels, like infrared, or wired cables, and take up different codes to finally be decoded, processed and displayed by a software system in the shape of a visual output. We suggested in Table 1 that artificial languages might be one of the possible modes involved, but the haptic channel actually constitutes a part of a complex human-machine interaction process.

Within the multimodal space of the game screen, the same concept can be represented by two or more synonymous signs simultaneously, each one belonging to a different representational system:

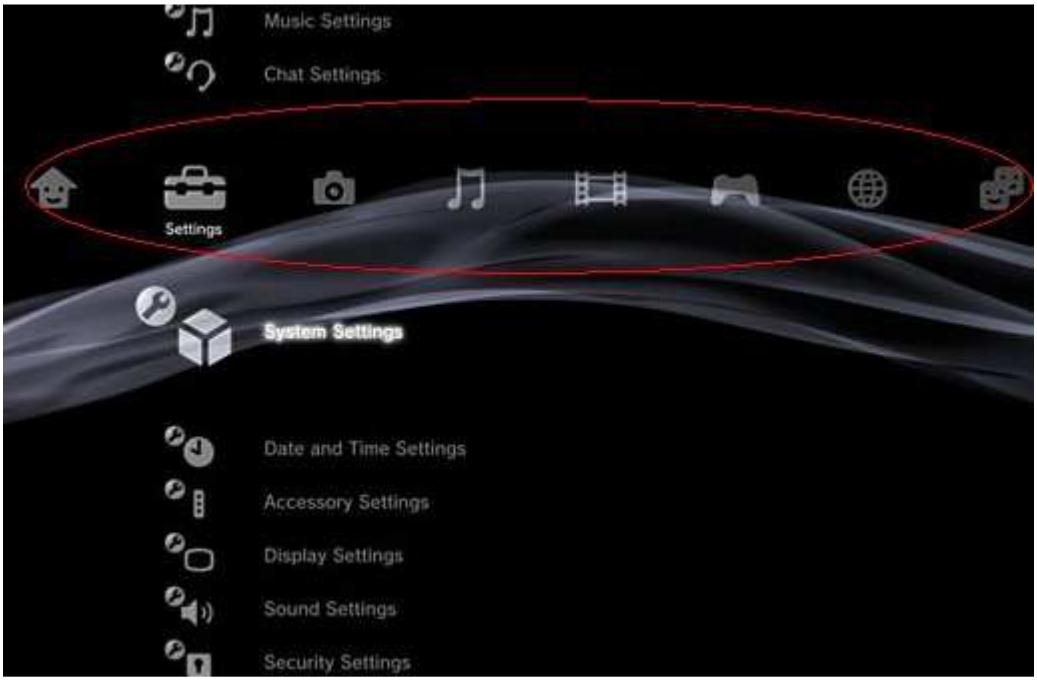

Figure 1: Cross Media Bar (XMB) Screen - Settings

This screenshot illustrates the multimedia tools activated by the console PlayStation 3. At the top of the screen, the horizontal bar of floating icons is highlighted by a circle, showing more or less familiar stylized symbols, similar to conventional clipboard or web site icons. From left to right, we can recognise - respectively - the Home, Settings, Photo, Music, Video, Game, Network and Chat menu icons. When the user scrolls through and selects them, each icon unfolds, displaying the corresponding written representation ("Settings" in this example) and the corresponding sub-items vertically, both in the iconic and written modes. In this communicative instance, the same concept can therefore be represented by two synonymous designations: the first is iconic - user-friendly but not transparent to users who are not familiar with gaming - and the second, linguistic. Further, the selection of screen options comes with sounds, therefore adding a third simultaneous representational system of a sensorial-acoustic kind to the same concept. Selection sounds, however, can be considered ambiguous representations resulting from the interaction with a concept: the user responds to a visual (iconic/written) concept (input) with a haptic output that is translated by the system into a sound.

In this example, technical concepts are inscribed in a representational system, the complexity of which can be assessed not only on a multimodal, but also on dynamic grounds. While nouns are static

(Settings), the synonymous icons can be considered interactive hyperlinks to a subset of concepts or submenu items (in this case System Settings, Date and Time Settings etc.). As a consequence, a terminological systematization becomes critical for managing multimodal technical concepts and giving instructions to the end user. Since the textual display of terminological units is not static, and not just verbal, terminological records need to define, classify and retrieve non-verbal units by means of the linguistic system traditionally used in terminological description, as well as include dynamic content/context.

3 MULTIMODAL OBJECTS

If we go into greater depth in the multimodal dissection of the technical concepts contained in entertainment software, we find out that the single modes sketched in Table 1 tend to recur in different combinations and meaningful patterns. Media studies scholars have recognised six typical multimodal compounds, or semiotic objects, recurring in entertainment software: graphics, sound, interface, gameplay, story and cut-scenes (adapted from Newman, 2005: 11). For the purposes of the present description, we will only highlight the three multimodal objects that are mostly relevant for specialised communication:

1 Graphics - Any images that are displayed and any effects performed on them. This includes 3D objects, 2D tiles, 2D full-screen shots, statistics, information overlays and anything else the player will see. (ibid.)

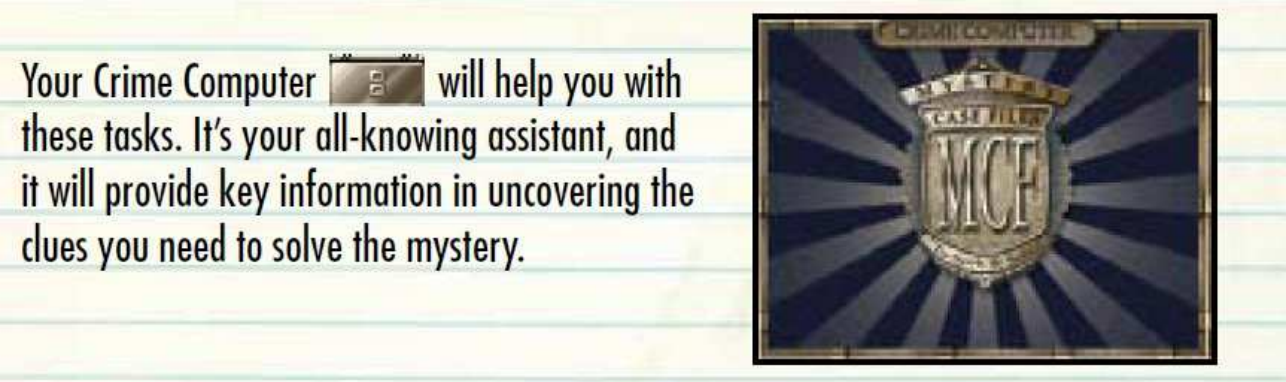

Figure 2: Game manual

This extract is drawn from the PDF manual of the game Mystery Case Files, MillionHeir (2008) for the console Nintendo DS. Due to the informative-illustrative function of game manuals, we can note that three modes exist and overlap in the representation of the same concept. The Crime Computer, a feature of the game, is represented through an icon, a verbal sign and illustrated by a screenshot, where the screenshot actually shows the Crime Computer in-game visual representation, i.e. the referent. In particular, iconic units have been shown to pose major challenges for a terminological classification within an ordered list of linguistic and/or iconic designations and because of the logical correspondence to the concept/caption they refer to (Soffritti, Bertaccini and Cortesi, 1997).

It is interesting to note that the linguistic code appears at two different levels of communication, or - we could say - in two distinct semiotic frames: the first one is the language of the manual and the second one is the linguistic signs contained within the visual system of the screenshot, where the written code has more of a graphic function than a linguistic one. In this context, we could assume that linguistic signs have a "parasitic function" within the frame of the visuals, as Roland Barthes argued with reference to photographic messages:

> "the text constitutes a parasitic message designed to connote the image, to 'quicken' it with one or more second-order signifiers. In other words, and this is an important historical reversal, the image no longer illustrates the words; it is now the words which, structurally, are parasitic on the image" (Barthes, 1981: 529).

2 Sound - "Any music or sound effects (SFX) that are played during the game. This includes starting music, CD music, MIDI, MOD tracks, Foley effects6, environmental sound" (Newman, 2005: 11).

Sounds may represent specific concepts according to the audio representational system coded by the software, like error or selection sounds, although they do not always specify one single referent/action.

The third relevant multimodal compound can be identified in interface objects, and is a symptomatic indicator of the revolution of textual supports as well as of the cognitive shift in the conceptualisation of interactive elements.

3 Interface - "anything that the player has to use or have direct contact with in order to play the game. It goes beyond simply the mouse/keyboard/joystick and includes graphics that the player must click on, menu systems that the player must navigate through and game control systems such as how to steer or control pieces in the game". (ibid.)

For example, the screenshot below displays a number of interactive iconic units and is another possible configuration of the PlayStation 3 Cross Media Bar, as referred to in Figure 1. From Figure 1, the player has scrolled horizontally through the main media bar icons (Home, Settings, Photo, Music, Video, Game, Network, PlayStation Network and Chat), selecting the Network label to the right, therefore activating a different textual configuration and the vertical unfolding of the Network menu.

Figure 3: Cross Media Bar (XMB) Screen – Interface elements

FOR AN INCLUSION OF NONVERBAL DYNAMIC UNITS IN TERMINOLOGY DESCRIPTION

In this discussion we have foregrounded multimodal elements and issues which are relevant to the specialised communication of interactive content, and need new modes of inclusion within terminological records, traditionally based on verbal description. One of the main emerging issues is certainly the electronic, non-verbal and dynamic nature of conceptual units in new textual supports, the definition and classification of which is however mostly limited to the use of the linguistic code. A second major shift can be identified in the conceptualisation strategies that combine abstract concepts with interaction, therefore incorporating a dynamic conceptualisation pattern on the part of the user.

Considering the relevance of iconic units within the specialised communication of entertainment software, one of the main problems is their labelling and retrieval within an ordered list of terminological records traditionally conceived for linguistic items. Their treatment as simple synonymous units may not be the most appropriate solution for a number of reasons. First, according to the standard ISO 704, synonymous designations do not constitute a definition. Paragraph 4.7 of ISO 704 does not recognise tautologies as being valid definitions. In theory, iconic conceptual units require a separate denotation and labelling - they could be indexed by means of alphanumeric characters, for instance - as well as a hierarchical classification consistent with the way in which the entertainment software system embeds hyperonimic or hyponimic iconic units (for example, in Figure 3, the first level icon Network, then Internet Search >

Download Management > Information Board etc. on a second level, and the respective subordinate icons on a third/fourth level).

Furthermore, iconic units serving the digital function of interface objects involve a definition which may not correspond to their expected linguistic label. For example, the context in Figure 1 and 3 clearly shows that icons are interactive and represent specific nomina actionis rather than abstract concepts. The icon of the networked world in Figure 3 not only represents a stylised (and slightly arbitrary or culture-bound) image of the "world/ Network" concept, it also stands for the action "Go to/Select the Network Menu", determining dynamically a different page configuration and an implicit conceptual set of operations in the mind of the user. As noted in previous research, the signs on computer screens do not necessarily refer to objects of the real world, but "carry information about complicated processes, and the signs may change form while the processes go on" (Järvi, 1997: 64).

Interactive icons therefore pose problems of a definitional nature because they combine a conceptual static denotation as well as a performative aspect, and their context is a moving multimodal texture of representational systems often coming with sounds. The linguistic definition of the iconic designation, as well as its labelling and contextual exemplification needs to account for such peculiarities. Help files including instructions to the user and terminological records could also provide dynamic sequences of context where the direct referent is a virtual onscreen action.

CONCLUDING REMARKS

Beyond fictional elements, video game functionality emerges as a form of LSP, just like the language of utility software, and conveys multimodal dynamic forms of conceptual representation and operationalisation. The requirements of modern digital communication problematize traditional semiotic models developed for the benefit of terminology theory and practice in relation to current LSP approaches primarily based on the verbal system. For the present and future state of the art, it seems that terminological issues need to be re-addressed in relation to new LSP communication needs, especially insofar as technical designations (linguistic and iconic) play a crucial role in the user's response and in the corporate policies of manufacturers: "Many authors have pointed to the need for 'integrative' approaches to LSP, that is, from system features to text features and further on to contextual and extra-verbal features. The increased interest in pragmatic and contextual aspects of language is a significant trend of linguistics in general, not only LSP research" (Myking, 2001: 48).

Finally, the epistemological and cognitive tendencies of modern communication, especially in the youngest generations, cannot be underestimated. Media studies have converged in emphasizing the diachronic impact of media crossover and hybridisation in human communication, since technologies have been shown to affect sensory perceptions, cognitive modes - and therefore conceptualisation patterns - of cultures through history (McLuhan, 1962; 1964), in terms of "visual turn" in the 19th century (Barthes, 1964), or "hyperreality" state (Baudrillard, 1981). In addition, the concept of "multimodal literacy" (Kress and Jewitt, 2003) emerges nowadays as a major topic in education, since multimodality pervades the cognitive experience of students (power point presentations, e-learning, web sites, video clips, CDs/DVDs, edutainment), no less than in their private lives. Since terminology serves the purposes of LSP in professional environments, it ought to take up the multimodal challenge as a key area for serving specialised fields in educational contexts. Such theories could also affect current views on sign models and conceptualisation patterns, insofar as virtual dynamic objects become the privileged referent of computer mediated communication (CMC) and human-machine interaction (HMI).

[1]According to the PricewaterhouseCoopers publication Global Entertainment and Media Outlook: 2008-2012, the video game market is one of the fastest growing sectors among the other media segments, with a turnaround worth 41.9 billion $ in 2007, and a compound annual growth rate of 10.3 %. (figures reported by Bond, 2008).
[2]Pac-Man, released in 1981, is one of the most famous crossover games, and a popular vintage title nowadays. The main character is a yellow dot eating other dots within a maze (Loguidice and Barton, 2009: 188). Sometimes images are more familiar than names:

[3]For a terminological discussion of these designations, see Bernal, 2006.
[4]According to the traditional model of communication elaborated by Jakobson, the channel is "physical or psychological connection between the addresser and addressee, enabling both of them to enter and stay in communication" (1987: 66).
[5]In the framework of multimodal theory, the term mode refers to "a regularised organised set of resources for meaning- making, including, image, gaze, gesture, movement, music, speech and sound-effect. Modes are broadly understood to be the effect of the work of culture in shaping material into resources for representation". (Jewitt and Kress, 2003:1).
[6]Foley effects are pre-recorded natural sounds (like a pair of scissors) added in the postproduction phase.

REFERENCES

BARTHES, R. (1964). Rhétorique de l´image. Communications, 4 1964, 40-51.

BARTHES, R. (1981). The Photographic Message, in Goldberg, V. (ed.) Photography in Print, Albuquerque: University of New Mexico Press, 521-533.

BAUDRILLARD, J. (1988). Simulacres et simulation. Paris: Galilée.

BERNAL, MERINO M. (2006). On the translation of videogames. JoSTrans, 6 2006, http://www.jostrans.org/issue06/art_bernal.php (consulted 02/09/2010).

BERTACCINI, F., CORTESI, C. and SOFFRITTI, M. (1997). L'icone dans la fiche terminologique: un nouveau point de départ?.Terminologie Nouvelles, 17 1997. Bruxelles, pp. 43 – 48.

BOND, J. (2008). Video game sales on winning streak, study projects. Reuters, http://www.reuters.com/article/technologyNews/idUSN1840038320080618 (consulted 13/09/2010)

BUDIN, G. (1997). Theoretical and Operational Problems of Semiotic Models in Terminology Theory. Terminology Science & Research, 8 (1/2) 1997, 79–83.

ECO, U. (1983). Travels in Hyperreality: Essays. W. Weaver (trans.). San Diego: Harcourt Brace Jovanovich.

ECO, U. (1984). Semiotica e filosofia del linguaggio. Torino: Einaudi.

ENSSLIN, A. (2010) Black and White: Language ideologies in computer game discourse, in Johnson S. and Milani T. (eds.) Language Ideologies and Media Discourse: Texts, Practices, Policies. London: Continuum, 205-222.

FRASCA, G. (2001). Videogames of the Oppressed: Videogames as a Means for Critical Thinking and Debate, Masters Thesis, http://www.ludology.org/articles/thesis/gamesandvideogames.html (consulted 13/03/2010)

GOTTI, M. (1991). I linguaggi specialistici: caratteristiche linguistiche e criteri pragmatici. Firenze: La Nuova Italia.

GOTTI, M. (2003). Specialized Discourse: Linguistic Features and Changing Conventions. Bern: Peter Lang.

JEWITT, C. and KRESS, G. (2003). Multimodal Literacy. New York: Peter Lang.

JAKOBSON, R. (1987). Language in Literature. Cambridge: Harvard University Press.

KOCOUREC, R (1982). La langue française de la technique et de la science. Wiesbaden: Oscar Brandstetter Verlag.

LOGUIDICE, B. and BARTON, M. (2009). Vintage Games: An Insider Look at the History of Grand Theft Auto, Super Mario, and the Most Influential Games of All Time. Oxford: Focal Press.

McLUHAN, M. (1962). The Gutenberg Galaxy. London: Routledge & Kegan Paul.

McLUHAN, M. (1964). Understanding Media. Canada: Mentor.

MYKING, J. (2001). Sign Models in terminology: Tendencies and Functions, LSP & Professional Communication, 1(2) (2001), 45-61.

NEWMAN, J. (2005). Videogames. London: Routledge.

JÄRVI, O. (1997). The Sign Theories of Eugen Wüster and Charles S. Peirce as Tools in Research of Graphical Computer User Interfaces. Terminology Science & Research, 8 (1/2) 1997, 63–72.

ROWSELL, J. (2006). Documenting Multimodal Literacies, http://www.gse.rutgers.edu/ContentScripts/genFile~FileFieldName~WordFormat~ContentItemID~res_2136~TableName~vwResources~MimeType~application%2Fmsword~VersionNumber~1.asp (consulted 13/09/2010)

SALEN, K. and ZIMMERMAN, E. (2006). Preface, in Salen, K. and Zimmerman, E. (eds.) The Game Design Reader: A Rules of Play Anthology, Massachusetts: MIT Press, XVI-XXVI.

SANDRINI, P. (1997). From Scratch to the Web: Terminological Theses at the University of Innsbruck. Terminology Science & Research, 8 (1/2) 1997, 127-136.

SHORTIS, T. (2001). The language of ICT: Information and communication technology. London: Routledge.

STÖCKL, H. (2004). In between modes: language and image in printed media, in E. Ventola, C. Charles, M. Kaltenbacher (eds.), Perspectives on Multimodality. Amsterdam /Philadelphia: John Benjamins, 9-30.

WÜSTER, E. (1979). Einführung in die allgemeine Terminologielehre und terminologische Lexikographie. Wien: Springer.